# Tweet Classification to Assist Human Moderation for Suicide Prevention

**Ramit Sawhney,**[*1] **Harshit Joshi,**[*2] **Alicia Nobles,**[*3] **Rajiv Ratn Shah,**[4]
[1] Netaji Subhas Institute of Technology
[2] University of Delhi
[3] University of California San Diego
[4] MIDAS, IIIT-Delhi
ramits.co@nsit.net.in, harshit113@ducic.ac.in, alnobles@health.ucsd.edu, rajivratn@iiitd.ac.in

## Abstract

Social media platforms are already engaged in leveraging existing online socio-technical systems to employ just-in-time interventions for suicide prevention to the public. These efforts primarily rely on self-reports of potential self-harm content that is reviewed by moderators. Most recently, platforms have employed automated models to identify self-harm content, but acknowledge that these automated models still struggle to understand the nuance of human language (e.g., sarcasm). By explicitly focusing on Twitter posts that could easily be misidentified by a model as expressing suicidal intent (i.e., they contain similar phrases such as "wanting to die"), our work examines the temporal differences in historical expressions of general and emotional language prior to a clear expression of suicidal intent. Additionally, we analyze time-aware neural models that build on these language variants and factors in the historical, emotional spectrum of a user's tweeting activity. The strongest model achieves high (statistically significant) performance (macro F1=0.804, recall=0.813) to identify social media indicative of suicidal intent. Using three use cases of tweets with phrases common to suicidal intent, we qualitatively analyze and interpret how such models decided if suicidal intent was present and discuss how these analyses may be used to alleviate the burden on human moderators within the known constraints of how moderation is performed (e.g., no access to the user's timeline). Finally, we discuss the ethical implications of such data-driven models and inferences about suicidal intent from social media. **Content warning: this article discusses self-harm and suicide.**

## Introduction

Suicide is a devastating public health issue resulting in nearly 800,000 lives lost worldwide annually (WHO 2014). Suicide prevention research and practice has made great strides in identifying general attributes of who is at higher risk for suicide (May and Klonsky 2016; Nock et al. 2012) and developing face-to-face psychosocial interventions (Hawton et al. 2016). However, despite this progress and growing awareness, suicide rates mirror rates from the 1950s (CDC 2020). Moreover, 80% of patients do not undergo clinical treatment, and, of those who do, about 60% of those who died of suicide denied having any suicidal thoughts to mental health practitioners (McHugh et al. 2019). In contrast, people exhibiting suicidal ideation often turn to social media to express their feelings (Coppersmith et al. 2018; Robinson et al. 2016; De Choudhury et al. 2016, 2013), with eight out of ten people disclosing their suicidal thoughts and plans on social media (Golden, Weiland, and Peterson 2009; Maple et al. 2019).

Public health experts and healthcare providers have called for innovative, real-time interventions, including leveraging digital technology like online interventions (Witt et al. 2017). Social media providers have responded to this call by introducing interventions for social media users who post about self-harm, including suicidal intent. For example, on Facebook and Twitter, users can flag posts that are potentially indicative of suicidal intent for trained human moderators to review. In turn, these moderators identify which posts warrant an intervention such as linking the at-risk user with support resources like the Substance Abuse and Mental Health Services Administration's National Suicide Prevention Lifeline (Twitter 2020; Card 2018)

Social media platforms are expanding these self-report mechanisms to include automatic identification of content suggestive of potential self-harm (Card 2018). Researchers have proved the feasibility of automatic identification of self-harm content (De Choudhury et al. 2016; Braithwaite et al. 2016). However, algorithms struggle with word sense disambiguation (i.e., when an individual may express similar phrases but differ semantically) that are seemingly easy for humans to identify (Card 2018; Sawhney et al. 2018b). For example, an algorithm may struggle to distinguish a sarcastic statement like *"Ugh. This class is so boring. I want to kill myself."* from an expression of suicidal intent. We refer to these types of tweets, where suicidal intent is absent but phrases that may be indicative of suicidal intent (e.g. "kill myself" in the previous example) are present, as edge cases.

Another limitation of recent advances for computationally assessing suicide risk on social media (Coppersmith et al. 2018; Ji et al. 2019) is that analyzing the linguistic traits of an individual tweet is often not sufficient for accurate identification of suicidal intent. Additional user-level contexts, such as posting history, can be instrumental in identifying a build-up of negative emotions, often linked to suicidal ideation (Robins et al. 1959; Sawhney et al. 2021). This

[*] Equal contribution

emotional build-up can occur weeks, months, or even years before the onset of suicidal ideation (Overholser 2003), and suicidal behavior is also influenced by past ideation or attempts (Van Heeringen and Marušic 2003).

Content moderation on social media remains a sociotechnical system requiring human labor to stay in the loop to rubber-stamp, or verify, automated decisions (Wagner 2019; Roberts 2019). Moderators report that the tasks can be not only "mind-numbingly repetitive", but also expose workers to emotional labor (Roberts 2019). Compounding this, the recent COVID-19 pandemic added to worker strain (Faddoul 2020), while simultaneously there was an increase in population mental health problems (Ayers et al. 2020) potentially increasing the burden on the few workers remaining.

We aim to further the task of automating identification of suicidal intent on social media, specifically Twitter, by differentiating between tweets containing potentially concerning language indicative of suicidal intent versus edge cases by examining the following research questions.

**RQ1: Are there temporal variations in linguistic features that differentiate between tweets containing expressions of true suicidal intent and language that could be misidentified as suicidal intent (i.e., edge cases)?** Using a Sparse Additive Generative Model (SAGE), we analyze how a user's language in their tweets varies temporally, differentiating between tweets where expressions of suicidal intent is present versus those where suicidal intent is absent.

**RQ2: Are there temporal variations in emotional language that differentiate between tweets containing expressions of true suicidal intent and edge cases?** We fine-tune a pre-trained transformer language model for emotions, and use it to automatically extract the differentiating temporal variations in emotions expressed in tweets where expressions of suicidal intent is present versus those where suicidal intent is absent.

**RQ3: Can predictive models, trained jointly on temporal activity and language features, differentiate between tweets containing expressions of true suicidal intent and edge cases?** We build a time-aware sequential neural model that differentiates between tweets where expressions of suicidal intent is present versus those where suicidal intent is absent. We then examine the interpretability of the model's decision on three example tweets.

**Findings and Contributions.** We find temporal variations in general and emotional language that are indicative of suicidal intent versus edge cases. We contribute a time-aware neural model that can differentiate if suicidal intent is present or absent in a tweet based on these language variations, by factoring in the relative time difference between successive historical tweets. We analyze how time-aware and emotion-aware models outperform previously published models of handcrafted language feature based approaches and recent deep learning neural network models to detect the presence of suicidal intent (i.e., we rebuilt these previously published models as well as trained and tested on an existing large dataset).

Specifically, we leverage STATENet, that advances the Bidirectional Encoder Representations from Transformers (BERT), a generic langauge model, by fine-tuning BERT based on Plutchik's wheel of emotions and apply a time-aware long short term memory model (LSTM) to model irregularities in a user's online behavior thereby incorporating historical context for each user. We show how time- and emotion-aware neural models advancs the state-of-the-art in user context models, specifically for suicidal intent detection by incorporating a well established emotion hierarchy, the Plutchik's wheel and by modeling the irregularities in user's tweeting behavior as opposed to conventionally used sequential models like RNNs, and LSTMs.

Finally, we discuss practical implications of such an interpretable model for human moderators including explainability, transparency, and moderator fatigue.

**Ethical Considerations and Privacy.** First, our study discusses self-harm and suicide, so we suggest caution for readers. Second, although our study was exempted from review from our respective institutional review boards, we adhere to suggested data protections (Benton, Coppersmith, and Dredze 2017; Chancellor et al. 2019) including separately storing the annotation of user data from raw user data on protected servers linked only through anonymous IDs. Third, although these tweets are public (Mishra et al. 2019; Fiesler and Proferes 2018), we paraphrase all example tweets per the moderate disguise scheme suggested by (Bruckman 2002) to avoid reverse identification (Ayers et al. 2018). Finally, our work focuses on developing a neural model for screening tweet content indicative of suicidal intent and does not make any diagnostic claims. We conclude by discussing ethical and privacy considerations in detail.

## Background

### Identifying Suicidal Content in Social Media

Previous research has established that our language and behavior on social media contain clues that indicate suicidal ideation offline. Although numerous studies have examined the detection of suicidal ideation on social media, we highlight research that delineates the various approaches to represent this task's complexity.

The earliest research focused on distinguishing tweets indicative of suicidal ideation versus non-suicidal ideation without considering the user's previous language or behavior (Braithwaite et al. 2016; O'Dea et al. 2015). Other researchers examined the group-level posts of people who shifted from posting on general mental health subreddits (e.g., depression) to a subreddit specifically for people actively contemplating suicide finding evidence that linguistic cues (e.g., a transition to more action-oriented words and self-references) are predictive of a switch between forums (De Choudhury et al. 2016). At the individual-user level, researchers have explored timelines of Weibo (Chinese social media) users with confirmed deaths by suicide finding evidence suggestive (increased frequency of posting and negative sentiment) of shifting temporal patterns in content and behavior prior to the user's death (Huang et al. 2017).

Outside of ongoing improvements to the proprietary algorithms developed in-house by social media platforms (Card 2018), there has been limited research on delineating tweets containing suicidal intent from edge cases, especially meth-

ods that incorporate temporal changes in the user's language and behavior. Our work aims to fill this gap examining the temporal attributes that delineate between these tweets.

### Human Moderation of Social Media

Human moderators are the unspoken gatekeepers of social media, underpinning moderation regardless of claims of automation (Wagner 2019). Of the two types of moderators: unpaid volunteer moderators (e.g., individuals flagging concerning content or subreddit moderators) and commercial content moderators (CCM) (e.g., employees or contractors of social media companies); we focus on CCMs.

Often, with little training, CCMs are faced with reading thousands of posts each day, flagged by concerned viewers (Weber and Seetharaman 2017), and are expected to elevate posts indicative of self-harm to experts specialized in rapid response (Goggin 2019). CCMs are often outsourced, isolating them from other workers, but ensuring cultural competency with the intended audience (Roberts 2019). This isolation is particularly troubling given their often relatively low status in the company, low wages, and frequent exposure to disturbing content (Roberts 2016).

CCMs report that some content is harder to adjudicate because of the effort that is required (e.g., viewing an entire video to determine if the content is contained within) (Roberts 2016), but also because they often are not given access to the user's profile to aid their decision of whether to elevate to the rapid response worker (Goggin 2019). The rapid response workers are encouraged to review a user's history for contextualization before making a decision, but little is known about this process, in part because it is proprietary. Current sociotechnical systems inherently have a human-in-the-loop despite automation and our work aims to build interpretable models that can be used by both CCMs and rapid response workers to aid in delineating the presence of true suicidal intent.

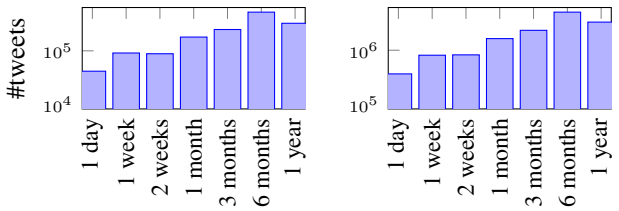
## Data

We build on an existing dataset of tweets annotated for the presence of suicidal intent that was curated by Sinha et al. (Sinha et al. 2019). First, we describe this existing dataset and, second, how we expanded this dataset to include each user's historical tweets. The existing dataset includes 34,306 tweets from 32,558 unique users that were collected using the Twitter REST API[1] from October 2018 to December 2018 and contained at least one of 248 suicidal phrases (e.g., "wanting to die", "suicide times", "last day", "feel pain point"). These suicidal phrases were identified by ranking phrases in subreddits containing discussions of suicide such as *r/suicidalthoughts*, *r/suicidewatch* by employing Pytextrank[2]. Under the supervision of a professional clinical psychologist, the tweets were annotated for the presence of suicidal intent by two students of clinical psychology as:

- **Suicidal Intent (SI) Present**: Tweets which explicitly describe suicidal ideation including plans, or previous attempts are discussed in a somber and non-flippant tone.

[1]https://developer.twitter.com/en/docs/twitter-api

[2]https://pypi.org/project/pytextrank/




(a) SI Present (b) SI Absent

Figure 1: Number of tweets in each time bucket $d$

- **Suicidal Intent (SI) Absent**: Tweets with no reasonable evidence for risk of suicide, including song lyrics, condolence messages, news and awareness-related tweets. In other words, consistent with our previous language, these are edge case tweets.

The annotators achieved a Cohen's Kappa score of 0.72. The resulting dataset contains 3,984 tweets annotated as SI present and 30,322 annotated as SI absent. For the disagreements, the two clinical psychology student annotators discussed the data sample and then it was further reviewed by the expert, practicing clinical psychologist who finalized the annotations for the disagreements. Further details of how this existing dataset was curated, including the qualitative labeling, can be found in (Mishra et al. 2019).

Our study builds on the previously described existing dataset by collecting all historical tweets for each of the 32,588 unique users in the existing dataset that were tweeted from 2009 to 2019, totalling to 2,314,127 historical tweets. The mean number of historical tweets per a user is 748 tweets (min = 0, max = 3,200) with a standard deviation of 789 tweets with 4,070 users having no historical tweets outside of the tweet in the existing data. The mean time difference between two consecutive tweets for a user is two days with a standard deviation of approximately 24 days between two tweets, indicative of large variations across users.

## RQ1: Exploring Temporal Linguistic Features of Tweets with SI Present versus SI Absent

Our first research question explores a user's historical tweets prior to the tweet in question to examine the temporal variations in linguistic features between tweets with SI present versus SI absent.

### Methods

We divide a user's historical tweets into non-overlapping time buckets prior to the tweet in question. The time buckets are defined in $d$ days, where $d \in \{1, 7, 14, \cdots, 365\}$ before the tweet in question, as shown in Figure 1. We then assess which words are distinctive of each time bucket using an unsupervised topic modeling technique, the Sparse Additive Generative Model (SAGE)[3] (Eisenstein, Ahmed, and Xing 2011). SAGE utilizes the measure of a log-odds ratio to contrast word distributions between a corpus of interest (in

[3]https://github.com/jacobeisenstein/SAGE

| Suicidal Intent Present | | Suicidal Intent Absent | |
|---|---|---|---|
| **1 Day** | **SAGE** | **1 Day** | **SAGE** |
| slit | 2.91 | dispatch | 2.48 |
| needles | 2.78 | neverland | 2.44 |
| schizophrenia | 2.42 | runaways | 2.16 |
| antidepressants | 2.23 | lobbying | 2.11 |
| urges | 2.13 | shutdown | 2.05 |
| **1 Week** | **SAGE** | **1 Week** | **SAGE** |
| selfloathing | 2.78 | bandersnatch | 2.84 |
| symbols | 2.14 | braveheart | 2.77 |
| resigned | 2.13 | birdbox | 2.68 |
| miscarriage | 1.98 | copycat | 2.39 |
| storytelling | 1.71 | lmmfaooooo | 2.31 |
| **2 Weeks** | **SAGE** | **2 Weeks** | **SAGE** |
| cbd | 2.38 | hamper | 2.00 |
| merry | 1.46 | camels | 2.00 |
| hearts | 1.56 | glances | 1.90 |
| pharma | 1.26 | obscene | 1.90 |
| reflux | 1.12 | reindeer | 1.88 |
| **1 Month** | **SAGE** | **1 Month** | **SAGE** |
| shouted | 1.77 | remember | 2.38 |
| vanishing | 1.56 | dalton | 1.97 |
| grammy | 1.41 | swerved | 1.91 |
| poems | 1.18 | piracy | 1.90 |
| thanksgiving | 1.18 | grammy | 1.57 |
| **3 Months** | **SAGE** | **3 Months** | **SAGE** |
| raffle | 1.85 | coveted | 2.01 |
| homosexuality | 1.36 | piranha | 1.89 |
| meaningless | 1.06 | raffle | 1.87 |
| morphine | 1.04 | prizes | 1.74 |
| fires | 1.03 | cnns | 1.69 |
| **6 Months** | **SAGE** | **6 Months** | **SAGE** |
| opines | 1.21 | iridescence | 1.26 |
| sweetner | 1.02 | constable | 1.13 |
| ford | 0.81 | bussiness | 1.10 |
| nike | 0.81 | fords | 1.06 |
| accusation | 0.79 | sweetener | 1.04 |
| **1 Year** | **SAGE** | **1 Year** | **SAGE** |
| laurel | 1.58 | rozza | 1.82 |
| sunglasses | 1.47 | shipping | 1.80 |
| rockets | 1.20 | got7s | 1.62 |
| celtics | 1.18 | laurel | 1.58 |
| autism | 1.15 | sunglasses | 1.48 |

Table 1: Five cherry-picked distinctive words across time-buckets obtained using SAGE for historic tweets prior to the tweet in question. A higher SAGE score is indicative of its saliency.

our case, the time bucket) against a baseline corpus (in our case, the whole corpus). SAGE focuses on identifying distinctive high-frequency terms with less sensitivity towards low-frequency words and identifies words that are more distinctive of the time bucket relative to all other time buckets. Prior to applying SAGE, we pre-processed the tweets by removing all hashtags and excluding tokens that appeared at less than five times in the corpus.

## Results

Table 1 presents the top five most distinctive words of historical tweets prior to the tweet in question (i.e., SI present versus SI absent) for each time bucket. In historical tweets one day prior to the tweet in question, words that are more frequently discussed in the context of mental health ("slit", "needles" and "schizophrenia") are more prominent in tweets where SI is present, especially in comparison to the neutral words prominent in tweets where SI is absent. For example,

> "· · · want a motionless body", "wrists **slit**", "· · · stop f**king with feelings", "good night fr* * * its over · · · "

In historical tweets one to two weeks prior to the tweet in question, words ("selfloathing", "resigned" and "miscarriage") commonly used to express distress are more prominent in tweets where SI is present and notably lacking from tweets where SI is absent. For example,

> "**selfloathing**", "hating my** · · · ", "yearned someone thin**", "prettier · · · ", "cleverer", "· · · and better"

Tweets older than two weeks include more seemingly random words ("grammy", "raffle", "nike"), but still on occasion include words reflective of groups at higher risk of suicide ("homosexuality", "autism") (Saha et al. 2019; Mayes et al. 2013) and "morphine" which can be reflective of chronic pain and substance use disorders, both of which are associated with higher suicide rates (Oquendo and Volkow 2018). Also, these historic tweets have more overlapping words across time buckets as suggested by lower SAGE scores and observe some overlapping words ("sunglasses") between the SI present and SI absent tweets.

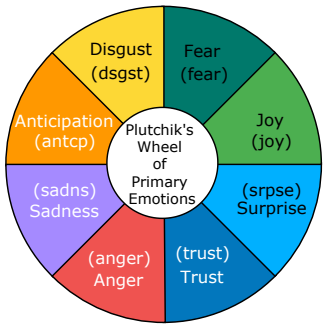



Figure 2: The primary set of opposing emotions described by Plutchik's wheel. In brackets: abbreviation code.

## RQ2: Exploring Temporal Emotional Language of Tweets with SI Present versus SI Absent

Our second research question explores a user's historical tweets prior to the tweet in question to examine the temporal variations in emotions expressed between tweets with SI present versus SI absent

## Methods

To analyze the emotional traits of a tweet, we utilize Plutchik's wheel of emotions (Plutchik 1980), which outlines eight primary emotions arranged as four pairs of opposing dualities, as shown in Figure 2. Building on existing

encoding methods in the state-of-the-art, STATENet (Sawhney et al. 2020), we utilize transfer learning to fine-tune a pre-trained BERT on Emonet (Mohammad et al. 2018), and call it EmotionBERT. Emonet is a popular corpus of 790,059 tweets labeled across the eight primary emotions. BERT fine-tuned on Emonet serves as an emotion feature extractor over historical tweets. We also tried extracting LIWC features from tweets to obtain representations of the tweets to feed to the predictive model, as opposed to the BERT encoder. We empirically found BERT to outperform LIWC significantly ($p < 0.05$) under identical conditions. We postulate this performance gap to BERT's ability to understand context owing to 1) massive pretraining and self-attention mechanisms that makes BERT a strong benchmark across many NLP problems. 2) fine-tuning BERT (Emotion BERT) captures a broader sense of both emotions and Twitter slang such as *"lmao"*, "lol", etc. better. After identifying the emotions in each historical tweet, for consecutive historical tweets, we first compute mean cosine similarity for emotions expressed. We then bin the tweets into the time buckets as previously described to examine how emotions vary temporally prior to the tweet in question. To determine statistical significance, we perform Wilcoxon Signed Rank Test (Woolson 2007) on emotions expressed in historical tweets between where SI is present and SI is absent.

### Results

We observe that the cosine similarity between emotions expressed by consecutive historic tweets associated with a tweet with SI present is significantly ($p < 0.001$) lower than a SI absent tweet. Figure 3 presents the variation in emotions identified in historical tweets prior to the tweet in question where SI is present. We observe that historical tweets where SI is present, include emotional variability that changes temporally, including less anticipation-related language three months prior, more sadness-related language two weeks prior, less joy-related language one week prior, and less fear-related language the day prior to the tweet in question. The difference in emotional variation is larger (i.e., language becomes more or less frequent for each emotion compared to the baseline where SI is absent) closer to the tweet in question, as shown by the intensity of the colors in Figure 3. The difference in emotions between tweets where SI is present versus absent reduces as the color intensity decreases with white, indicating no difference in the emotional variation.

## RQ3: Predictive Modeling of Tweets with SI Present versus SI Absent

Our first two research questions focused on exploring temporal language features. Our final question focuses on building and interpreting a predictive model based on these exploratory results to identify tweets where SI is present.

### Methods

**Problem Formulation** Following prior work, we formulate the problem as a binary classification task to predict a label $y_i$ for a tweet in question $t_i$ authored by user $u_i$, where, $y_i \in \{SI\ present,\ SI\ absent\}$. We denote a tweet to be assessed for the presence of suicidal intent as $t_i \in T = \{t_1, t_2, \cdots, t_N\}$ authored by a user $u_j \in U = \{u_1, u_2, \cdots, u_M\}$, posted at time $\tau^i_{curr}$. Each tweet $t_i$ is associated with history $H_i = [(h^i_1, \tau^i_1), (h^i_2, \tau^i_2), \cdots, (h^i_L, \tau^i_L)]$ where $h^i_k$ is a historic tweet by the user $u_j$ posted at time $\tau^i_k$ with $\tau^i_1 <) \tau^i_2 < \cdots < \tau^i_L < \tau^i_{curr}$. We now describe the model architecture (Figure 4), training, and performance assessment.

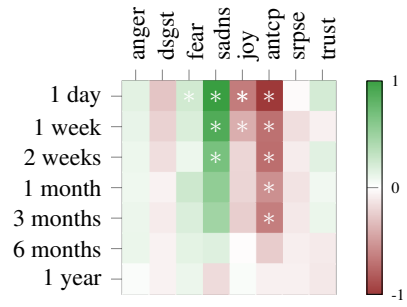



Figure 3: Temporal variation in the eight primary emotions expressed for historical tweets prior to the tweet in question (here we visualize a tweet where SI is present). Green: higher emotional value for tweets where SI is present than where SI is absent, Red: lower emotional value, White: no change. $*$ denotes statistical significance ($p < 0.005$) under Wilcoxon's Signed Rank Test.

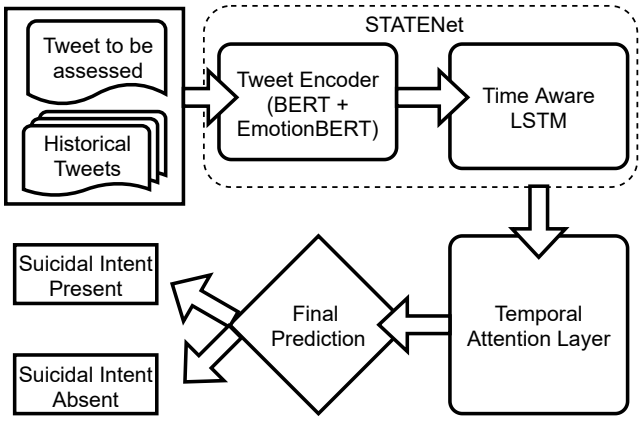



Figure 4: Model Overview

**Extracting General Language Features from Tweets** Recent developments in NLP have shown that pre-trained transformer models, like BERT, yield more comprehensive representations of linguistic features in a tweet (Salminen et al. 2020). We use BERT to encode linguistic representations of tweets by averaging the output vectors for all tokens in each tweet from BERT's final layer. Formally, $T_i^{'} = \text{BERT}(t_i)$ where $T_i^{'} \in \mathbb{R}^{768}$ is linearly transformed using a dense layer to $T_i \in \mathbb{R}^d$ with dimension $d$.

**Extracting Emotional Language Features from Tweets** We use Emotion BERT (as discussed in RQ2), which takes a historical tweet $h_k^i$ and a high-dimensional encoding is obtained from the `[CLS]` token from the final transformer layer which is densely connected with a 8-dimensional output layer representative of each primary emotion. We utilize the latent 768-dimensional encoding as the representation of the emotional spectrum. Formally, we define the emotion vector ($E_k^i \in \mathbb{R}^{768}$) of each historic tweet as:

$$E_k^i = \texttt{EmotionBERT}(h_k^i) \tag{1}$$

**Learning Temporal Tweeting Activity Patterns** We focus on the state-of-the-art, STATENet (Sawhney et al. 2020) for our subsequent model analysis. The sequential nature of a user's tweets over time makes sequential models, like LSTM models, the most natural methods to encode representations and learn from a sequence of a user's historical tweets. LSTM cells assume that the input is equally spaced sequences and thus are unable to model irregularities in posting times of tweets. However, tweets occur at irregular time intervals that can vary widely from a few seconds to a few years (Wojcik and Hughes 2019) and capturing these temporal variations may in itself be important for characterizing changes in behavior.

To overcome this limitation, STATENet leverages a Time-aware LSTM (T-LSTM) (Baytas et al. 2017), where the T-LSTM cell incorporates the actual time differences between tweets as well as each tweet's extracted features $T_k^i$ as previously described. Using this relative time difference between the user's historical tweets can progressively model the user's language and tweeting activity more accurately over time. The T-LSTM applies time decay to the memory according to the elapsed time between successive tweets and weights the short-term memory cell $C_k^S$. This translates to the intuitive understanding that the greater the time elapsed between two tweets, the less impact they should have on each other. For each historic tweet $h_k^i$, the T-LSTM cell modifies LSTM gate operations to compute the current hidden state ($\tilde{H}_k^i \in \mathbb{R}^d$) by feeding an altered memory cell.

**Temporal Attention** Often only a few tweets contain relevant signals. For example, (Shing, Resnik, and Oard 2020) report that for a user showing signs of suicidal intent, experts identified only two out of 1,326 postings were relevant to assess the presence of suicidal intent. To capture such relevant signals, we propose a temporal attention mechanism. This mechanism learns adaptive weights for contextual representations of each tweet ($\tilde{H}_k^i$), highlighting tweets with indicative markers for the presence of suicidal intent and aggregates them as:

$$a_i = \sum_{k=1}^{L_i} \alpha_k^i \tilde{H}_k^i, \; \alpha_k^i = \frac{exp(\tilde{\alpha}_k^i)}{\sum_{k=1}^{L_i} exp(\tilde{\alpha}_k^i)} \tag{2}$$

$$\tilde{\alpha}_k^i = c^L tanh(W_x \tilde{H}_k^i + b_x) \tag{3}$$

where $W_x \in \mathbb{R}^{T_i \times H}$, $b_x \in \mathbb{R}^{T_i}$ and $c \in \mathbb{R}^{T_i}$ are network parameters and $a_i$ is the contextual representation of a user's historical tweets.

**Training the Predictive Model** To identify the presence (or absence) of SI in a tweet in question, these models jointly learn from the language of the tweet in question and the temporal variation in emotion in historical tweets. We feed extracted features $a_i$ from the Temporal Attention to a dense layer with Rectified Linear Unit ($ReLU$) to form a prediction vector. Finally, a softmax function is used to output the probability of whether the tweet in question has SI present.

$$\tilde{y}_i = ReLU(W_y(a_i) + b_y) \tag{4}$$

$$\hat{y}_i = softmax(\tilde{y}_i) \tag{5}$$

where $\hat{y}_i$ is the final suicide risk assessment and $\{W_y, b_y\}$ are network parameters.

Suicide has a low baseline rate in the general population and subsequently tweets with true SI present are relatively small (Ji et al. 2019). To address this problem of class imbalance, which is likely much higher in reality than our dataset, we train the neural network using Class-Balanced loss (Cui et al. 2019) along with Focal Loss (Lin et al. 2017). This loss function applies a class-wise re-weighting scheme by introducing a weighting factor that is inversely proportional to the number of samples.

**Data Split and Preprocessing** We split the tweets in the dataset on the basis of users such that there is no overlap between users in the train, validation, and test set. We perform a stratified 70:10:20 split across the three sets, such that the train, validation, and test sets consist of 24014, 3431, and 6861 tweets, respectively. Although there may be multiple tweets to be assessed by the same user, their associated history differs according to the tweets' posting timestamps. We ensure that for each tweet to be assessed, only the historical tweets having timestamps older than that of the tweet to be assessed are used for historic modeling.

**Experimental Setup and Hyperparameters** We select hyperparameters based on the highest Macro F1 obtained on the validation set for all models. We use grid search to explore: number of features in hidden state $\tilde{H}^D \in \{8, 64, \cdots, 512\}$, number of LSTM layers $n \in \{1, 2, 5\}$, dropout $\delta \in \{0.0, 0.1, \cdots, 0.5\}$, initial learning rate $I_{lr} \in \{0.01, 0.001, 0.0001\}$, warm-up steps $S_{ws} \in \{3, 5, 7\}$. The optimal hyperparameters were found to be: $\tilde{H}^D = 512$, $n = 1$, $\delta = 0.5$, $I_{lr} = 0.0001$, $S_{ws} = 5$. We implement all methods with PyTorch 1.6 (Paszke et al. 2019) and optimize using mini-batch Adam with a batch size of 256 and $I_{lr} = 0.0001$. We use the cosine scheduler with a warmup step of 5 (Gotmare et al. 2018). We train the model for 20 epochs and apply early stopping with a patience of 5 epochs. The model takes 934s to train on an Nvidia Tesla K80 GPU.

**Performance Comparisons** We compare the performance of these state-of-the-art methods via replications of the architectures and representations presented in prior works on detection of suicidal content in social media:

1. Random Forest (Sawhney et al. 2018b): We replicate a random forest with tweet level features including statistical, LIWC features, n-grams and part of speech counts.

| Contextual Modeling | Language Features | Model | Macro F1 ↑ | Precision ↑ | Recall ↑ | Accuracy ↑ |
|---|---|---|---|---|---|---|
| No History | LIWC+POS+TF-IDF | Random Forest | 0.536 | 0.489 | 0.513 | 0.548 |
| | POS+Interaction+Content | Logistic Regression | 0.571 | 0.563 | 0.583 | 0.589 |
| | Word2Vec | C-LSTM | 0.588 | 0.568 | 0.597 | 0.602 |
| History + RNN | Fast Text Embeddings | SDM | 0.743 | 0.578 | 0.755 | 0.819 |
| | BERT Embeddings | DualContextBert | 0.767 | 0.589 | 0.786 | 0.823 |
| Specific Temporal Functions | BERT Embeddings | Exponential Decay | 0.737 | 0.582 | 0.759 | 0.828 |
| | | Surprise and Episodic Modeling | 0.741 | 0.583 | 0.762 | 0.831 |
| **History+Time-LSTM** | BERT Embeddings | **STATENet + Temporal Attention** | **0.804*** | **0.612*** | **0.813*** | **0.856*** |

Table 2: Median of metrics for various representations and architectures obtained over 20 different runs. LIWC = Linguistic Inquiry and Word Count, POS = part of speech, TF-IDF = term frequency–inverse document frequency. * indicates that the result is significantly better than DualContextBert ($p < 0.005$) for Wilcoxon Signed Rank Test. **Bold** denotes best performance.

2. Logistic Regression (De Choudhury et al. 2016): We replicate a logicstic regression classifier that utilizes unigram and bigrams tokens with linguistic structure, interpersonal awareness and interaction based measure categories as features.
3. C-LSTM (Sawhney et al. 2018a): We replicate the deep neural network that uses CNN to capture local features and LSTMs for tweet encoding.
4. Suicide Detection Model (SDM) (Cao et al. 2019): We replicate a model that encodes tweets using fine-tuned FastText embeddings. Historic tweets were passed sequentially through LSTM + attention and concatenated with the tweet to be assessed.
5. Exponential Decay (Sinha et al. 2019): We replicate a model that weighs GloVe embeddings of historic tweets through an exponential decay and ensembles it with the GloVe embedding trained on a BiLSTM + attention.
6. Surprise and Episodic Modeling (Mathur et al. 2020): We replicate the a decision level ensemble model similar to exponential decay, but factors in sinusoidal and white Gaussian noise for historic tweet modeling.
7. DualContextBert (Matero et al. 2019): We replicate the best performing model at CLPsych 2019, which uses BERT for encoding Reddit posts fed to an attention-based RNN layer. In our implementation, we use all the user's historic tweets.

**Interpreting the Model** We present visualizations that could be supplied to human moderators to support interpretation of the model's final decision of whether SI is present or absent in the tweet in question. These visualizations show the salient information the model used for its decision without revealing the user's profile, much like the conditions described for CCMs.

## Results

**Performance Comparison** Table 2 presents the results of STATENet with existing models, indicating that STATENet is significantly ($p < 0.005$) outperforming these previous architectures and representations. As shown in Table 2, sequential models outperform those relying on text-based features alone. Additionally, models which factor in exact time lapses and applies attention mechanism to identify tweets that show signs of suicidal intent, outperforms other models that model tweets as equally spaced inputs. We postulate that time-aware sequential models likely learn more accurate representations of users' historic emotional state (Sawhney et al. 2021).

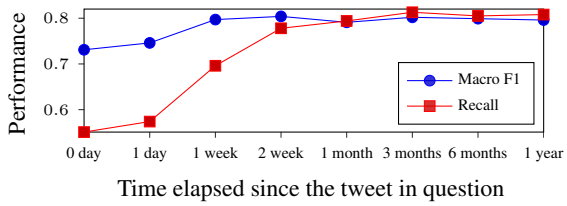



Figure 5: Performance with number of historical tweets

We observe low precision of models since tweets showing suicidal intent form only a tiny proportion of the data. However, posts on social media are copious, and a high recall for SI present can help prioritize at-risk users. Hence, achieving a higher recall can be crucial in a human-centered mental health model and thereby balance low precision.

**Results with Varying User History** We discuss model performance given a different amount of historical lookback in terms of number of days in Figure 5. We can see that Macro F1 and Recall for SI present monotonically increase until 2 weeks, likely as the model learns more about the user's emotional state over time. When we further increase the number of tweets that the model can see in user history, the model performance does not significantly improve. This observation is consistent with previously published research (Selby, Yen, and Spirito 2013; Kaplow et al. 2014; Glenn et al. 2020) finding that the importance of a user's emotional state decreases over time for assessing the user's current suicide risk. Additionally, we note that short-term behavior upto 1 month (as shown in Figure 1a) is more bursty whereas longer term behavior shows greater density linking towards slow temporal variation (Sawhney et al. 2021).

**Model Interpretation** Figure 6 presents three interesting example tweets in question as they could be presented to a human moderator and the correct and incorrect assessment made by different models for the presented tweets. We show

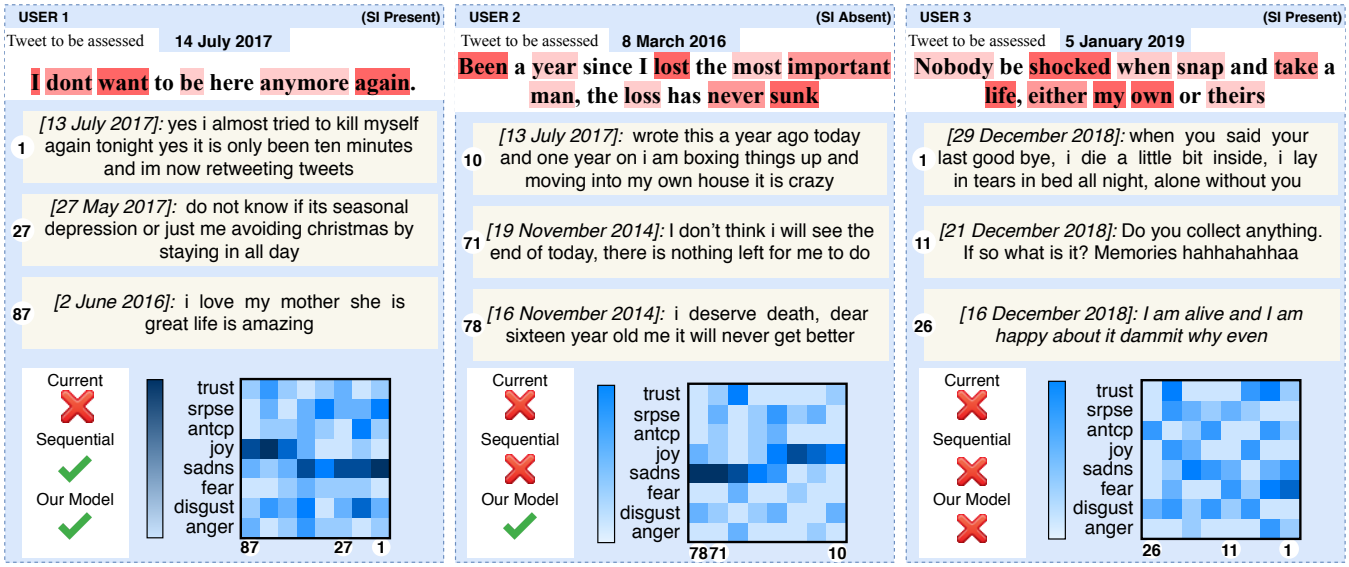



Figure 6: Tweet to be assessed and chronologically ordered historic user tweets. Visualized self-attention (averaged over all 12 BERT heads, higher red intensity corresponds to higher attention). Figure adapted from (Sawhney et al. 2020)

the same tweets as used in (Sawhney et al. 2020).[4] For User 1, relying on examining the tweet in question itself is too ambiguous to assess if the user has suicidal intent. However, in Figure 6, the text of the historical tweets clearly indicate a history of suicidal behavior and the emotional intensity indicates a history of tweets with reduced joy-related language and increased sadness-related language. In contrast, consider User 2, where the model has associated the word "loss" with the presence of SI. The user historically posted ambiguous tweets that may indicate SI in years prior as shown in Figure 6, but the user has reduced their sadness-related language and increased their joy-related language most recently. For User 3, the current tweet does not contain strong semantic indicators of suicidal intent. Moreover, historic tweets do not show any recognizable emotional pattern. Such a case presents the complexities associated with suicide risk assessment and exemplifies the complexity of the task and need for future work.

## Discussion

Social media platforms are already leveraging online sociotechnical systems to provide just-in-time interventions. Our work expands on pre-existing work that exemplified that suicidal content can be automatically identified to examine how language features may contribute to improved identification and provide interpretable insights that can aid human moderation in identifying expressions of suicidal intent in difficult tweets that include language potentially indicative of suicidal intent. We elaborate on the influence of temporal context in improving automatic identification of suicidal intent and implications for the role that interpretable models can play in human moderation of suicidal intent.

[4]https://github.com/midas-research/STATENet_Time_Aware_Suicide_Assessment

### Importance of Temporal Context in Evaluating Suicidal Intent of Social Media Posts

Our results suggest that there is temporal variation in general and emotional language before a clear expression of suicidal intent and consideration of this variation can greatly improve performance of automatic identification of suicidal intent. For example, consistent with (Glenn et al. 2020; Huang et al. 2017), we find shifting patterns in emotional language, including increased language indicative of fear and sadness and decreased language indicative of anticipation and joy, within two weeks of the tweet containing the presence of suicidal intent. In comparison to edge case tweets that do not contain suicidal intent, emotions are less variable (i.e., historical posts of clear displays of suicidal intent do not switch between joy and sadness, but instead consistently increase usage of emotional language indicative of sadness). In offline settings, amplification of emotional factors such as emotional reactivity (Tarrier et al. 2007), intensity (Links et al. 2008), and instability (Palmier-Claus et al. 2012) are known predecessors before suicidal behavior. Moreover, consideration of how temporally close social media posts are can lead to a model learning the relative importance of timing, which is important given that social media users often have irregular posting behaviors ranging from seconds to years (Wojcik and Hughes 2019).

### Implications for Human Moderation

Moderators are subjected to repetitive, quota-driven queues of content in difficult working conditions, but little is known about the strategies or tools they use to employ decision-making, in part because this information is proprietary and many sign a non-disclosure agreement (Roberts 2019). Much of the research to date regarding human moderation has focused on the emotional labor and working conditions of moderators (Gray and Suri 2019), how moderators

strategize moderation rules and engage with their communities (more relevant for volunteer governance) (Seering et al. 2019), and how users perceive moderated content (Jhaver et al. 2019). In light of this, we draw parallels from a high stress workplace with many notifications and high risk decision making that has been well studied - the intensive care unit (ICU). We acknowledge the great differences in these workplaces, but see many parallels to moderation.

"Alarm fatigue" or when alarms are so excessive, many of which are false positives, that healthcare providers become desensitized from alarms has been well documented and studied in the ICU (Drew et al. 2014). A number of prediction algorithms and visualization tools have been developed to help with alarm fatigue (Joshi et al. 2016). We also point to another issue with alarm fatigue. That is, there is a trade-off between precision-recall in these environments to avoid alarm fatigue (Drew et al. 2014). Similar trade-offs should be explored when developing algorithms that identify mental health-related problems in social media (Eskisabel-Azpiazu, Cerezo-Menéndez, and Gayo-Avello 2017). For example, one moderator reported saving nine lives out of 800 videos flagged for suicidal intent (Roberts 2019). If a moderator is constantly faced with false positives, similar to providers in the ICU, moderators may also become fatigued and wary of the accuracy of new notifications. A data-driven algorithm that balances false alarms and points to specific evidence of the user's temporal language and behavior may help filter the sheer volume of content one must review to come to a decision as well as reduce the emotional labor involved with reviewing content. However, we acknowledge that health-related inferences based on social media versus healthcare system data are different and less is currently known about the ethical and liability considerations for precision and recall trade-offs in this domain.

## Ethical, Privacy, and Transparency Considerations

With growing concerns about privacy and data ownership, we would be remiss to not mention the potential implications of just-in-time interventions outside of the clinical context.

First, in the clinical context, data protections exist like those under the US Health Insurance Portability and Accountability Act of 1996 (HIPAA) exist. Social media companies have not been transparent in the inferences made about a user's mental health, how the data used to support those inferences or the inferences themselves are stored, or the third-party companies they may work with for moderation (Goggin 2019). This is a serious concern given the growing number of security breaches, unethical data management (Cadwalladr and Graham-Harrison 2018), and stigmatization of mental health. As a matter of fact, the European Union has halted Facebook's algorithms because they collect sensitive health information (Goggin 2019).

Second, many social media users intend for their posts to only be read by their targeted audience and are largely unaware that their data may be used for other purposes (Fiesler and Proferes 2018). Additionally, many think sensitive data about their mental health should be similarly protected much like clinical data under HIPPA (Andalibi and Buss 2020). In turn, social media platforms leave it to the individual users to navigate the jargon-filled terms of service to figure out what data is being collected and how to manage their privacy settings (Roberts 2019). Third, users may belong to vulnerable groups (e.g., over-policed or individuals without documentation) and prefer to toggle their settings to avoid unnecessary contact with law enforcement (i.e., wellness checks). Similarly, false positives increase unnecessary contact with law enforcement. For example, false positives have led to police requiring a person who was not suicidal to go to the hospital for a mental health evaluation and police leaking personal information to news media (Goggin 2019).

Finally, as (Eskisabel-Azpiazu, Cerezo-Menéndez, and Gayo-Avello 2017) and (Nobles et al. 2018) mention, there is an open ethical question about who should have access to the outcomes of this moderation (e.g., the individual, their friends, their family, mental health professionals). Some researchers have circumvented this by allowing their participants to opt-in and ensuring secure data storage in accordance with HIPPA requirements (Eskisabel-Azpiazu, Cerezo-Menéndez, and Gayo-Avello 2017). This is in contrast to the current, non-transparent data collection and inference that social media platforms are performing. Moreover, in other contexts, social media users have expressed confusion and frustration at the lack of transparency in moderation (Jhaver et al. 2019). Similarly, users may have interest in why their tweet was deemed to be at risk and an interpretable algorithm may be a step towards providing explanations.

## Limitations and Future Work

First, we acknowledge that interpretation of expressions of suicidal intent is subjective, variable across social media users, and we do not know the true outcomes of the user behind the post. Future work, such as (Nobles et al. 2018) does with text messages, could explore the integration of social media posts with known outcomes. Second, we acknowledge that even the best model is not perfect with 20% of posts still misidentified; in practice a false alarm rate this high is still debilitating to a human moderator and, importantly, could result in the loss of human life. Future work should focus on examining features that can further improve performance and lead to interpretable models. We acknowledge that models and humans are often a source of bias; interpretable models can help identify these biases. Finally, our work has hypothesized that interpretable models can aid human moderators in identifying true expressions of suicidal intent in social media posts. Future work should confirm if interpretable models do indeed reduce burden and help human moderators. Utilizing words obtained from the SAGE analysis as features for logistic regression can highlight the significant difference in the words of the two sets, SI Present and SI Absent, and forms our future work.

## Acknowledgements

Dr. Nobles was supported by NIH NIDA K25 DA049944.

## References

Andalibi, N.; and Buss, J. 2020. The Human in Emotion Recognition on Social Media: Attitudes, Outcomes, Risks.

In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16.

Ayers, J. W.; Caputi, T. L.; Nebeker, C.; and Dredze, M. 2018. Don't quote me: reverse identification of research participants in social media studies. *NPJ Digital Medicine* 1(30). doi:10.1038/s41746-018-0036-2.

Ayers, J. W.; Leas, E. C.; Johnson, D. C.; Poliak, A.; Althouse, B. M.; Dredze, M.; and Nobles, A. L. 2020. Internet Searches for Acute Anxiety During the Early Stages of the COVID-19 Pandemic. *JAMA Internal Medicine* ISSN 2168-6106. doi:10.1001/jamainternmed.2020.3305. URL https://doi.org/10.1001/jamainternmed.2020.3305.

Baytas, I. M.; Xiao, C.; Zhang, X.; Wang, F.; Jain, A. K.; and Zhou, J. 2017. Patient subtyping via time-aware LSTM networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, 65–74.

Benton, A.; Coppersmith, G.; and Dredze, M. 2017. Ethical Research Protocols for Social Media Health Research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 94–102. Valencia, Spain. doi:10.18653/v1/W17-1612. URL https://www.aclweb.org/anthology/W17-1612.

Braithwaite, S. R.; Giraud-Carrier, C.; West, J.; Barnes, M. D.; and Hanson, C. L. 2016. Validating Machine Learning Algorithms for Twitter Data Against Established Measures of Suicidality. *JMIR Mental Health* 3(2): e21. doi:10.2196/mental.4822. URL https://doi.org/10.2196/mental.4822.

Bruckman, A. 2002. Studying the amateur artist: A perspective on disguising data collected in human subjects research on the Internet. *Ethics and Information Technology* 4(3): 217–231.

Cadwalladr, C.; and Graham-Harrison, E. 2018. Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The guardian* 17: 22.

Cao, L.; Zhang, H.; Feng, L.; Wei, Z.; Wang, X.; Li, N.; and He, X. 2019. Latent Suicide Risk Detection on Microblog via Suicide-Oriented Word Embeddings and Layered Attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 1718–1728.

Card, C. 2018. How Facebook AI helps suicide prevention. *Facebook Newsroom* URL https://about.fb.com/news/2018/09/inside-feed-suicide-prevention-and-ai/.

CDC. 2020. WISQARS™ — Web-based Injury Statistics Query and Reporting System. https://www.cdc.gov/injury/wisqars/index.html. Accessed: 2020-09-02.

Chancellor, S.; Birnbaum, M. L.; Caine, E. D.; Silenzio, V. M. B.; and De Choudhury, M. 2019. A Taxonomy of Ethical Tensions in Inferring Mental Health States from Social Media. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 79–88. New York, NY, USA. ISBN 9781450361255. doi:10.1145/3287560.3287587. URL https://doi.org/10.1145/3287560.3287587.

Coppersmith, G.; Leary, R.; Crutchley, P.; and Fine, A. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights* 10: 1178222618792860.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9268–9277.

De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2013. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.

De Choudhury, M.; Kiciman, E.; Dredze, M.; Coppersmith, G.; and Kumar, M. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, 2098–2110.

Drew, B. J.; Harris, P.; Zègre-Hemsey, J. K.; Mammone, T.; Schindler, D.; Salas-Boni, R.; Bai, Y.; Tinoco, A.; Ding, Q.; and Hu, X. 2014. Insights into the problem of alarm fatigue with physiologic monitor devices: a comprehensive observational study of consecutive intensive care unit patients. *PloS one* 9(10): e110274.

Eisenstein, J.; Ahmed, A.; and Xing, E. P. 2011. Sparse additive generative models of text. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, 1041–1048. Citeseer.

Eskisabel-Azpiazu, A.; Cerezo-Menéndez, R.; and Gayo-Avello, D. 2017. An ethical inquiry into youth suicide prevention using social media mining. *Internet Research Ethics for the Social Age* 227.

Faddoul, M. 2020. *COVID-19 is triggering a massive experiment in algorithmic content moderation*. URL https://www.brookings.edu/techstream/covid-19-is-triggering-a-massive-experiment-in-algorithmic-content-moderation/. (accessed August 26, 2020).

Fiesler, C.; and Proferes, N. 2018. "Participant" perceptions of Twitter research ethics. *Social Media+ Society* 4(1): 2056305118763366.

Glenn, J. J.; Nobles, A. L.; Barnes, L. E.; and Teachman, B. A. 2020. Can Text Messages Identify Suicide Risk in Real Time? A Within-Subjects Pilot Examination of Temporally Sensitive Markers of Suicide Risk. *Clinical Psychological Science* 8(4): 704–722.

Goggin, B. 2019. *Inside Facebook's suicide algorithm: Here's how the company uses artificial intelligence to predict your mental state from your posts*. URL https://www.businessinsider.com/facebook-is-using-ai-to-try-to-predict-if-youre-suicidal-2018-12. (accessed August 25, 2020).

Golden, R. N.; Weiland, C.; and Peterson, F. 2009. *The truth about illness and disease*. Infobase Publishing.

Gotmare, A.; Keskar, N. S.; Xiong, C.; and Socher, R. 2018. A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. *arXiv preprint arXiv:1810.13243* .

Gray, M. L.; and Suri, S. 2019. *Ghost work: how to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books.

Hawton, K.; Witt, K. G.; Salisbury, T. L. T.; Arensman, E.; Gunnell, D.; Hazell, P.; Townsend, E.; and van Heeringen, K. 2016. Psychosocial interventions following self-harm in adults: a systematic review and meta-analysis. *The Lancet Psychiatry* 3(8): 740–750.

Huang, X.; Xing, L.; Brubaker, J. R.; and Paul, M. J. 2017. Exploring Timelines of Confirmed Suicide Incidents Through Social Media. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, 470–477.

Jhaver, S.; Appling, D. S.; Gilbert, E.; and Bruckman, A. 2019. " Did You Suspect the Post Would be Removed?" Understanding User Reactions to Content Removals on Reddit. *Proceedings of the ACM on human-computer interaction* 3(CSCW): 1–33.

Ji, S.; Pan, S.; Li, X.; Cambria, E.; Long, G.; and Huang, Z. 2019. Suicidal Ideation Detection: A Review of Machine Learning Methods and Applications. *arXiv:1910.12611* .

Joshi, R.; van Pul, C.; Atallah, L.; Feijs, L.; Van Huffel, S.; and Andriessen, P. 2016. Pattern discovery in critical alarms originating from neonates under intensive care. *Physiological Measurement* 37(4): 564.

Kaplow, J. B.; Gipson, P. Y.; Horwitz, A. G.; Burch, B. N.; and King, C. A. 2014. Emotional suppression mediates the relation between adverse life events and adolescent suicide: Implications for prevention. *Prevention science* 15(2): 177–185.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Links, P. S.; Eynan, R.; Heisel, M. J.; and Nisenbaum, R. 2008. Elements of Affective Instability Associated with Suicidal Behaviour in Patients with Borderline Personality Disorder. *The Canadian Journal of Psychiatry* 53(2): 112–116. doi:10.1177/070674370805300206.

Maple, M.; Frey, L. M.; McKay, K.; Coker, S.; and Grey, S. 2019. "Nobody Hears a Silent Cry for Help": Suicide Attempt Survivors' Experiences of Disclosing During and After a Crisis. *Archives of suicide research* 1–19.

Matero, M.; Idnani, A.; Son, Y.; Giorgi, S.; Vu, H.; Zamani, M.; Limbachiya, P.; Guntuku, S. C.; and Schwartz, H. A. 2019. Suicide risk assessment with multi-level dual-context language and bert. In *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology*, 39–44.

Mathur, P.; Sawhney, R.; Chopra, S.; Leekha, M.; and Shah, R. R. 2020. Utilizing Temporal Psycholinguistic Cues for Suicidal Intent Estimation. In *Lecture Notes in Computer Science*, 265–271. Springer International Publishing. doi:10.1007/978-3-030-45442-5_33. URL https://doi.org/10.1007/978-3-030-45442-5_33.

May, A. M.; and Klonsky, E. D. 2016. What distinguishes suicide attempters from suicide ideators? A meta-analysis of potential factors. *Clinical Psychology: Science and Practice* 23(1): 5–20.

Mayes, S. D.; Gorman, A. A.; Hillwig-Garcia, J.; and Syed, E. 2013. Suicide ideation and attempts in children with autism. *Research in Autism Spectrum Disorders* 7(1): 109–119.

McHugh, C. M.; Corderoy, A.; Ryan, C. J.; Hickie, I. B.; and Large, M. M. 2019. Association between suicidal ideation and suicide: meta-analyses of odds ratios, sensitivity, specificity and positive predictive value. *BJPsych open* 5(2).

Mishra, R.; Prakhar Sinha, P.; Sawhney, R.; Mahata, D.; Mathur, P.; and Ratn Shah, R. 2019. SNAP-BATNET: Cascading Author Profiling and Social Network Graphs for Suicide Ideation Detection on Social Media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 147–156. Minneapolis, Minnesota. doi:10.18653/v1/N19-3019. URL https://www.aclweb.org/anthology/N19-3019.

Mohammad, S.; Bravo-Marquez, F.; Salameh, M.; and Kiritchenko, S. 2018. SemEval-2018 Task 1: Affect in Tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, 1–17. New Orleans, Louisiana. doi:10.18653/v1/S18-1001. URL https://www.aclweb.org/anthology/S18-1001.

Nobles, A. L.; Glenn, J. J.; Kowsari, K.; Teachman, B. A.; and Barnes, L. E. 2018. Identification of Imminent Suicide Risk Among Young Adults Using Text Messages. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, 1–11. New York, NY, USA. ISBN 9781450356206. doi:10.1145/3173574.3173987. URL https://doi.org/10.1145/3173574.3173987.

Nock, M.; Nock, M. K.; Borges, G.; and Ono, Y. 2012. *Suicide: Global perspectives from the WHO world mental health surveys*. Cambridge University Press.

O'Dea, B.; Wan, S.; Batterham, P. J.; Calear, A. L.; Paris, C.; and Christensen, H. 2015. Detecting suicidality on Twitter. *Internet Interventions* 2(2): 183 – 188. ISSN 2214-7829. doi:https://doi.org/10.1016/j.invent.2015.03.005. URL http://www.sciencedirect.com/science/article/pii/S2214782915000160.

Oquendo, M. A.; and Volkow, N. D. 2018. Suicide: a silent contributor to opioid-overdose deaths. *New England journal of medicine* 378(17): 1567–1569.

Overholser, J. 2003. Predisposing factors in suicide attempts: life stressors. In *Evaluating and treating adolescent suicide attempters*, 41–52. Elsevier.

Palmier-Claus, J.; Taylor, P.; Varese, F.; and Pratt, D. 2012. Does unstable mood increase risk of suicide? Theory, research and practice. *Journal of affective disorders* 143(1-3): 5–15.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch:

An Imperative Style, High-Performance Deep Learning Library. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32*, 8024–8035. Curran Associates, Inc. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Plutchik, R. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, 3–33. Elsevier.

Roberts, S. T. 2016. Commercial content moderation: Digital laborers' dirty work. *The Intersectional Internet: Race, Sex, Class and Culture Online* .

Roberts, S. T. 2019. *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.

Robins, E.; Murphy, G. E.; Wilkinson Jr, R. H.; Gassner, S.; and Kayes, J. 1959. Some clinical considerations in the prevention of suicide based on a study of 134 successful suicides. *American Journal of Public Health and the Nations Health* 49(7): 888–899.

Robinson, J.; Cox, G.; Bailey, E.; Hetrick, S.; Rodrigues, M.; Fisher, S.; and Herrman, H. 2016. Social media and suicide prevention: a systematic review. *Early intervention in psychiatry* 10(2): 103–121.

Saha, K.; Kim, S. C.; Reddy, M. D.; Carter, A. J.; Sharma, E.; Haimson, O. L.; and De Choudhury, M. 2019. The language of lgbtq+ minority stress experiences on social media. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW): 1–22.

Salminen, J.; Hopf, M.; Chowdhury, S. A.; Jung, S.-g.; Almerekhi, H.; and Jansen, B. J. 2020. Developing an online hate classifier for multiple social media platforms. *Human-centric Computing and Information Sciences* 10(1): 1. ISSN 2192-1962. doi:10.1186/s13673-019-0205-6. URL https://doi.org/10.1186/s13673-019-0205-6.

Sawhney, R.; Joshi, H.; Flek, L.; and Shah, R. R. 2021. PHASE: Learning Emotional Phase-aware Representations for Suicide Ideation Detection on Social Media. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*.

Sawhney, R.; Joshi, H.; Gandhi, S.; and Shah, R. R. 2020. A Time-Aware Transformer Based Model for Suicide Ideation Detection on Social Media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7685–7697. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.emnlp-main.619. URL https://www.aclweb.org/anthology/2020.emnlp-main.619.

Sawhney, R.; Manchanda, P.; Mathur, P.; Shah, R.; and Singh, R. 2018a. Exploring and learning suicidal ideation connotations on social media with deep learning. In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 167–175.

Sawhney, R.; Manchanda, P.; Singh, R.; and Aggarwal, S. 2018b. A Computational Approach to Feature Extraction for Identification of Suicidal Ideation in Tweets. In *Proceedings of ACL 2018, Student Research Workshop*, 91–98. Melbourne, Australia. doi:10.18653/v1/P18-3013. URL https://www.aclweb.org/anthology/P18-3013.

Seering, J.; Wang, T.; Yoon, J.; and Kaufman, G. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* 21(7): 1417–1443.

Selby, E. A.; Yen, S.; and Spirito, A. 2013. Time varying prediction of thoughts of death and suicidal ideation in adolescents: weekly ratings over 6-month follow-up. *Journal of Clinical Child & Adolescent Psychology* 42(4): 481–495.

Shing, H.-C.; Resnik, P.; and Oard, D. W. 2020. A Prioritization Model for Suicidality Risk Assessment. In *Proceedings of ACL 2020*, 8124–8137.

Sinha, P. P.; Mishra, R.; Sawhney, R.; Mahata, D.; Shah, R. R.; and Liu, H. 2019. # suicidal-A Multipronged Approach to Identify and Explore Suicidal Ideation in Twitter. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 941–950.

Tarrier, N.; Gooding, P.; Gregg, L.; Johnson, J.; and Drake, R. 2007. Suicide schema in schizophrenia: The effect of emotional reactivity, negative symptoms and schema elaboration. *Behaviour Research and Therapy* 45(9): 2090–2097. doi:10.1016/j.brat.2007.03.007. URL https://doi.org/10.1016/j.brat.2007.03.007.

Twitter, I. 2020. *About self-harm and suicide*. URL https://help.twitter.com/en/safety-and-security/self-harm-and-suicide. (accessed August 27 2020).

Van Heeringen, C.; and Marušic, A. 2003. Understanding the suicidal brain. *The British Journal of Psychiatry* 183(4): 282–284.

Wagner, B. 2019. Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems. *Policy & Internet* 11(1): 104–122. doi:10.1002/poi3.198. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/poi3.198.

Weber, L.; and Seetharaman, D. 2017. *The Worst Job in Technology: Staring at Human Depravity to Keep It Off Facebook*. URL https://www.wsj.com/articles/the-worst-job-in-technology-staring-at-human-depravity-to-keep-it-off-facebook-1514398398. (accessed August 25, 2020).

WHO. 2014. *Preventing suicide: A global imperative*. World Health Organization.

Witt, K.; Spittal, M. J.; Carter, G.; Pirkis, J.; Hetrick, S.; Currier, D.; Robinson, J.; and Milner, A. 2017. Effectiveness of online and mobile telephone applications ('apps') for the self-management of suicidal ideation and self-harm: a systematic review and meta-analysis. *BMC psychiatry* 17(1): 297.

Wojcik, S.; and Hughes, A. 2019. Sizing up Twitter users. https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/. Accessed: 2020-07-27.

Woolson, R. 2007. Wilcoxon signed-rank test. *Wiley encyclopedia of clinical trials* 1–3.